DOCUMENT CATEGORIZING METHOD, DOCUMENT CATEGORIZING APPARATUS, AND STORAGE MEDIUM ON WHICH A DOCUMENT CATEGORIZATION PROGRAM IS STORED

Inventors: Michihiro Nagaishi
Shinji Miwa

BACKGROUND OF THE INVENTION

Background Art

The present invention relates to a document categorizing method, a document categorizing apparatus, and a storage medium including a document categorizing program stored thereon, for categorizing a large number of documents into a plurality of clusters according to semantic similarity.

In the art of categorizing a large number of documents into a plurality of clusters according to semantic similarity, it is known to extract feature elements from the respective documents and categorize the documents according to the extracted feature elements. One specific categorization method is to employ the whole of each document (the entire content of each document including a title and a body) as a target and extract feature elements from the entire contents of the respective documents. After that, the documents are categorized into a plurality of clusters on the basis of the extracted feature elements.

In the case where feature elements are extracted from the entire contents of the respective documents, very complicated processing is needed in morphological analysis and extraction of feature elements. Thus, when such processing is performed by a central processing unit (CPU) in an information processing apparatus, a large load is imposed upon the CPU. Documents generally include many expressions having no direct relationships with the purport thereof. Therefore, if documents are categorized according to feature elements extracted by searching the entire contents of the respective documents, the resultant categorization is often meaningless. That is, a large number of noise clusters are created.

One technique to solve the above problem is to first extract a title representing the purport of a document, then extract a feature element from the title, and finally categorize the document on the basis of the extracted feature element.

It is thought that this technique allows documents to be correctly categorized according to the purport of the respective documents.

As described above, several techniques of categorizing documents into clusters are known.

However, even when documents are categorized into clusters on the basis of feature elements extracted from the titles of the documents, the number of resultant clusters often becomes too great for user to use the resultant information. For example, when a large number of clusters obtained as a result of categorization are compared with one another, many same documents can be included in different clusters. In such a case, a user has to search the large number of presented clusters to find desired information. This is very inconvenient for users.

In view of the above, it is an object of the present invention to provide a technique to recategorize a large number of categorized clusters into a simplified easily-understandable form by means of merging clusters.

Brief Description of the Drawings

Fig. 1 is a block diagram illustrating a first embodiment of the present invention.

Fig. 2 shows examples of documents for illustration of the first embodiment of the present invention.

Fig. 3 is a flow chart illustrating the outline of a document categorization process according to the first embodiment of the present invention.

Fig. 4 illustrates an example of the content of a feature table representing the relationship between feature elements and documents.

Fig. 5 illustrates a result obtained by categorizing the documents on the basis of the feature table shown in Fig. 4.

Fig. 6 illustrates a process of merging two clusters, wherein examples of documents included in the respective clusters are shown.

Fig. 7 illustrates a result obtained by performing a cluster merging process upon the categorization result shown in Fig. 5.

Fig. 8 is block diagram of a document categorizing apparatus which performs a cluster merging process in accordance with in what manner feature elements appear in original documents.

Fig. 9 is a block diagram illustrating a second embodiment of the present invention.

Fig. 10 shows examples of documents for illustration of the second embodiment of the present invention.

Fig. 11 is a flow chart illustrating the outline of a document categorization process according to the second embodiment of the present invention.

Fig. 12 illustrates an example of the content of a feature table representing the relationship between feature elements and documents.

Fig. 13 illustrates a result obtained by categorizing the documents on the basis of the feature table shown in Fig. 12.

Fig. 14 illustrates a process of merging two clusters, wherein examples of documents included in the respective clusters are shown.

Fig. 15 illustrates a result obtained by performing a cluster merging process upon the categorization result shown in Fig. 13.

Fig. 16 shows an example of a categorization result which is displayed such that the cluster names of clusters which have been merged into a final cluster are represented in an AND form (that is, the respective cluster names are placed in a single horizontal line).

Fig. 17 shows an example of a categorization result which is displayed such that the cluster names of clusters which have been merged into a final cluster are represented in another AND form (that is, the respective cluster names are placed in different lines).

Disclosure of the Invention

To achieve the object described above, the present invention provides a document categorizing method for categorizing a plurality of documents into a plurality of clusters according to semantic similarity, the method being characterized in that after categorizing the plurality of documents into a plurality of clusters according to semantic similarity, and a cluster merging process is performed such that relations among clusters of the plurality of clusters are evaluated on the basis of documents included in the respective clusters, and two or more clusters having a degree of relation equal to or higher than a predetermined value are combined together.

Preferably, the cluster merging process is performed such that the evaluation of relations among clusters under consideration as to whether they should be merged or not is performed on the basis of the number of documents commonly included in the clusters under consideration relative to the total number of documents included in the clusters under consideration, and cluster merging is performed in accordance with the evaluation result.

Alternatively, the cluster merging process may be performed such that in what manner feature elements, which characterize respective clusters under consideration as to whether they should be merged or not, appear in the respective clusters under consideration is examined, and cluster merging is performed in accordance with the manner in which the feature elements appear.

Preferably, The cluster merging process is performed at least for two clusters, and after completion of the cluster merging process a first time, the cluster merging process is performed repeatedly for the resultant set of clusters until no further cluster merging occurs.

Preferably, after completion of the cluster merging process, supplementary information indicating that cluster merging has been performed and also indicating the basis on which the cluster merging has been performed is output.

In the present invention, as described above, after categorizing documents into a plurality of clusters, the cluster merging process is performed such that relations among clusters of the plurality of clusters are evaluated on the basis of documents included in the respective clusters, and two or more clusters having a degree of relation equal to or higher than the predetermined value are combined together. Even when a large number of clusters have been generated in a first-time clustering process, the degrees of relations among the generated clusters are evaluated and clusters having high degrees of relations are combined together, and a simplified categorization result is presented to a user. This allows the user to find desired information in a highly efficient manner.

Because the evaluation of relations among clusters under consideration as to whether they should be merged or not is performed on the basis of the number of documents commonly included in the clusters under consideration relative to the total number of documents included in the clusters under consideration, the cluster merging process can be performed easily and correctly.

The evaluation of relations among clusters may be performed such that in what manner feature elements appear in the respective clusters under consideration as to whether they should be merged or not is examined, and cluster merging may be performed in accordance with the manner in which the feature elements appear. In this method, because the evaluation of the degree of relations among clusters is performed on the basis of the actual contents of documents, the cluster merging process can be performed in a more proper fashion.

The cluster merging process is performed for a combination of at least two clusters, and after completion of the cluster merging process, the cluster merging process is performed repeatedly for the set of clusters obtained in the previous cluster merging process until no further cluster merging occurs, thereby making it possible to obtain a simplified categorization result.

After completion of the cluster merging process, supplementary information indicating that cluster merging has been performed and also indicating the basis on which the cluster merging has been performed is output. Thus, a use can know in what manner the cluster merging process has been performed. This makes it possible for the user to use the supplementary information to find desired information from the result of the cluster merging process.

According to a second aspect of the present invention, there is provided a document categorizing method for categorizing a plurality of documents into a plurality of clusters according to semantic similarity, the method being characterized in that after categorizing the plurality of documents into a plurality of clusters according to semantic similarity, a cluster merging process is performed such that relations among clusters of the plurality of clusters are evaluated on the basis of documents included in the respective clusters, and two or more clusters having a degree of relation equal to or higher than a predetermined value are combined together; information representing which clusters have bee merged together and also representing the degrees of relation among the merged clusters is generated and the information is output together with the categorization result to be presented to a user so that when final clusters obtained as a result of the cluster merging process are displayed, the user can see in what manner the cluster merging process has been performed to obtain the final cluster.

Preferably, the information output so as to enable the user to see in what manner the cluster merging process has been performed is given by modifying the manner of displaying the cluster names of respective clusters merged together in accordance with the degree of relation among the clusters merged together in such a manner that when the degree of relation among the clusters is higher than a predetermined value, the cluster names are displayed in an AND form, however when the degree of relation among the clusters is lower than the predetermined value, the cluster names are displayed in an OR form.

Preferably, when the cluster names are displayed in the AND form, the cluster names of the respective clusters are displayed successively in a single horizontal line or the respective cluster names are displayed in different lines, while when the cluster names are displayed in the OR form, a delimiter is inserted between adjacent cluster names of the respective clusters.

When a certain cluster includes a cluster therein, the name of the cluster included in the certain cluster may be enclosed within brackets and placed after the name of the certain cluster.

In the present invention as described above, cluster-merging-process information is generated which represents which clusters have been merged together and also represents the degrees of relation among the merged clusters, and the cluster-merging-process information is displayed when final clusters obtained via the cluster merging process are displayed so that a user can see in what manner the cluster merging process has been performed to obtain the final cluster.

This makes it possible for the user to easily understand which clusters have been merged into which final clusters and can know the degrees of relations among the clusters merged together, simply by seeing the information displayed. The information output so as to enable the user to see what relations the clusters have is given by modifying the manner of displaying the cluster names of respective clusters merged together in accordance with the degrees of relations among the clusters merged together.

More specifically, when the degree of relation among clusters is higher than a predetermined value, the cluster names are displayed in an AND form, however when the degree of relation among the clusters is lower than the predetermined value, the cluster names are displayed in an OR form. For example, when the degree of relation is very high, the cluster names may be displayed successively in a single horizontal line or may be displayed in different lines such that one name is placed in one line. In the case where the degree of relation is not very high, a delimiter may be inserted between adjacent cluster names. When the user sees the cluster names displayed in one of the above manners, he/she can understand from which clusters the cluster has been created via the cluster merging process and can know the degree of the relation among the original clusters.

In the case where a certain cluster includes another cluster therein, the cluster name of the cluster included in the first cluster may be enclosed within brackets after the cluster name of the first cluster name. This allows the inclusive relation to be represented in a simple manner.

The present invention also provides a document categorizing apparatus for categorizing a plurality of documents into a plurality of clusters according to semantic similarity, the apparatus comprising a clustering unit for categorizing a plurality of documents into a plurality of clusters in accordance with semantic similarity, and a cluster merging unit which evaluates the relation among the plurality of clusters created by the clustering unit on the basis of the documents included in the respective clusters and then combines two or more clusters having a degree of relation equal to or higher than a predetermined value.

Furthermore, the present invention also provides another document categorizing apparatus for categorizing a plurality of documents into a plurality of clusters according to semantic similarity, the apparatus comprising a clustering unit for categorizing a plurality of documents into a plurality of clusters in accordance with semantic similarity, a cluster merging unit which evaluates the relation among the plurality of clusters created by the clustering unit on the basis of the documents included in the respective clusters and then combines two or more clusters having a degree of relation equal to or higher than a predetermined value, a cluster-merging-process information generator for generating cluster-merging-process information representing which clusters have been merged together and also representing the degrees of relation among the merged clusters wherein the cluster-merging-process information is to be displayed when final clusters obtained via the cluster merging process performed by the cluster merging unit are displayed so that a user can see in what manner the cluster merging process has been performed to obtain the final cluster, and categorization result outputting means for outputting the cluster-merging-process information such that the cluster-merging-process information is included in the categorization result to be presented to the user.

The present invention also provides a storage medium on which a document categorizing program for categorizing a plurality of documents into a plurality of clusters according to semantic similarity is stored, the document categorizing program comprising a clustering step for categorizing a plurality of documents into a plurality of clusters in accordance with semantic similarity, and a cluster merging step in which the degrees of relation among clusters of the plurality of clusters obtained in the clustering step are evaluated on the basis of documents included in the respective clusters, and two or more clusters having a degree of relation equal to or higher than a predetermined value are combined together.

Furthermore, the present invention provides another storage medium on which a document categorizing program for categorizing a plurality of documents into a plurality of clusters according to semantic similarity is stored, the document categorizing program comprising a clustering step for categorizing a plurality of documents into a plurality of clusters in accordance with semantic similarity, a cluster merging step in which the degrees of relation among clusters of the plurality of clusters obtained in the clustering step are evaluated on the basis of documents included in the respective clusters, and two or more clusters having a degree of relation equal to or higher than a predetermined value are combined together, a cluster-merging-process information generating step for generating cluster-merging-process information representing which clusters have been merged together and also representing the degrees of relation among the merged clusters wherein the cluster-merging-process information is to be displayed when final clusters obtained via the cluster merging process performed by the cluster merging unit are displayed so that a user can see in what manner the cluster merging process has been performed to obtain the final cluster, and a step for outputting the cluster-merging-process information such that the cluster-merging-process information is included in the categorization result to be presented to the user.

Best Mode for Carrying Out the Invention

(First Embodiment)

A first embodiment of the present invention is described below. Note that the technique described herein can be applied not only to a document categorizing method and a document categorization apparatus according to the present invention but also to a document categorization program stored on a storage medium according to the present invention.

In the present embodiment, titles are first extracted from respective documents, then feature elements are extracted from the titles, and finally the document is categorized according to the extracted feature elements.

Fig. 1 illustrates the structure of the apparatus according to the present embodiment. As shown in Fig. 1, the apparatus consists mainly of a clustering unit 1 for categorizing a plurality of documents into a plurality of clusters in accordance with semantic similarity, a cluster merging unit 2 which evaluates the relation among the plurality of clusters created by the clustering unit 1 on the basis of the documents included in the respective clusters and then combines two or more clusters having a degree of relation equal to or higher than a predetermined value, and a categorization result outputting unit 3 for outputting the categorization result obtained via the cluster merging process performed by the cluster merging unit 2.

The clustering unit 1 includes a document storage unit 11, a sentence analyzer 12, a feature element extractor 13, a feature table generator 14, a document categorizing unit 15, and a categorization result storage unit 16.

The cluster merging unit 2 serves to combine clusters, as will be described in detail later.

The categorization result outputting unit 3 includes an output control unit 31 and a display unit 32. The categorization result outputting unit 3 controls the process of outputting the result of the cluster merging process performed by the cluster merging unit 2.

The document storage unit 11 included in the clustering unit 1 stores a large number of document data in a database. Hereinafter, it is assumed that documents shown in Fig. 2 are categorized. The documents shown in Fig. 2 include different documents D1, D2,..., D7 including titles T2, T2,..., T7 and main bodies A1, A1,..., A7, respectively.

The sentence analyzer 12 analyzes the documents stored in the document storage unit 11 to extract the titles of the respective documents. The extraction of the titles is performed by the document analyzer 12 as follows.

A first method is to detect a part defined as a title according to the document format, and employ the detected part as a title if such a part is detected. A second method is to detect a part specified to be displayed with characters having a greater size than a standard size, and employ the detected part as a title if such a part is detected. A third method is to extract a predetermined number of sentences or words described at the beginning of a document and employ the extracted sentence or words as a title. The first, second, and third methods described above are performed sequentially. When the first method is performed, if a part defined as a title is detected, the detected part is employed as the title. If a part defined as a title is not detected, the second method is performed. If a part specified to be displayed with large-sized characters is detected, the detected part is employed as the title. If a part specified to be displayed with large-sized characters is not detected, the third method is performed to detect a title.

The feature element extractor 13 extracts feature elements from the respective document titles detected by the sentence analyzer 2.

The feature table generating means 14 produces a feature table representing the relationship between the feature elements detected from the titles and the respective documents. A specific example of the content of the feature table will be described later.

The document categorizing unit 15 examines the content of the feature table and categorizes the documents D1, D2,..., D7 into a plurality of clusters according to semantic similarity. More specifically, documents including a common feature element are detected on the basis of the feature elements included in the titles of the documents D1, D2,..., D7, and the detected documents are categorized into a cluster. The document categorization unit 15 includes a synonymous feature dictionary (not shown). Categorization into clusters may also be performed using the synonymous feature dictionary as follows. When documents including a common feature element are detected, a judgment as to whether a common feature element is included or not is made by determining whether a synonym is included or not using the synonymous feature dictionary. If synonyms are detected in documents, those documents are categorized into the same cluster.

The categorization result storage unit 16 stores the content obtained via the categorization performed by the document categorization unit 15.

The document categorization process performed by the apparatus constructed in the above-described manner according to the present invention is described below. In the document categorization process according to the present embodiment, as shown in the flow chart of Fig. 3, a large number of documents to be processed are first categorized into a plurality of clusters according to semantic similarity (step S1). Thereafter, relations among the clusters are evaluated on the basis of the documents included in the respective clusters as will be described in detail later (step S2). Two or more clusters having a degree of relation higher than a predetermined level are combined together (step S3). The document categorization process is described in further detail below with reference to a specific example.

Herein, it is assumed that the documents D1, D2,..., D7 shown in Fig. 2 are categorized. In the present embodiment, feature elements are extracted from the titles of the respective documents, and clustering is performed on the basis of the extracted feature elements. Thereafter, a cluster merging process is performed upon the result obtained through the above clustering process. First, the process of extracting feature elements from the titles and performing the clustering on the basis of the extracted feature elements (by the clustering unit 1) is described. The document analyzer 12 detects the titles of the respective documents D1, D2,..., D7. For example, the title T1 is detected from the document D1, the title T2 from the document D2, the title T3 from the document D3, and so on. Thus, the titles T1, T2,..., T7 are detected from the respective documents D1, D2,..., D7.

The feature element extractor 13 then extracts feature elements from the respective titles. Thereafter, the feature table generator 14 produces a feature table representing the relationships between the feature elements and the documents including the feature elements in their titles. An example of a feature table is shown in Fig. 4. In this example, the feature table represents the relationship between feature elements each included in three or more different documents and the documents including the feature elements. Numerals described in the feature table represent the numbers of feature elements included in the respective documents. For example, in the case of feature element "paper", one feature element is included in the title of each of the documents D1, D4, D6, and D7.

As can be seen from the feature table shown in Fig. 4, documents including "paper" as a feature element in their title are D1, D4, D6, and D7, documents including "cassette" as a feature element in their title are D1, D4, and D7, and documents including "installation" as a feature element in their title are D2, D3, D5 and D7. In Fig. 2, these feature elements are underlined.

The document analyzer 15 access the feature table and categorizes the respective documents into clusters for each feature element. The result of the categorization is shown in Fig. 5. In the categorization into clusters, as described earlier, the detection of feature elements commonly included in documents may be performed by detecting synonyms included in documents, using the synonymous feature dictionary, and documents including detected synonyms may be categorized into the same document cluster. For example, when "paper" and "printing paper" are extracted as feature elements, documents including either one of these feature elements are categorized into the same cluster.

The obtained categorization result is stored in the categorization result storage unit 16. In the categorization result shown in Fig. 5, as for a cluster categorized as "paper" (including documents D1, D4, D6, and D7), as can be understood from the document contents shown in Fig. 2, a paper cassette is described in the content of the document D1, setting of paper in the document D4, a smear created on printed paper in the document D6, and installation of an additional paper cassette in the document D7.

Thus, the categorization of the documents D1, D4, D6, and D7 including a description about paper into the cluster can be regarded as proper.

In the case of a cluster categorized as "cassette" (including documents D1, D4, and D7), as can be understood from the document contents shown in Fig. 2, a paper cassette is described in the content of the document D1, setting of paper in the document D4, and installation of an additional paper cassette in the document D7.

The contents of the documents D1, D4, D6 and D7 include a description about setting of paper, and thus the categorization of these documents into the cluster can be regarded as proper.

In the case of a cluster categorized as "installation" (including documents D2, D3, D5, and D7), as can be understood from the document contents shown in Fig. 2, installation of an additional memory is described in the content of the document D2, installation of an interface card in the document D3, installation of an additional hard disk in the document D5, and installation of an additional paper cassette is in the document D7.

The contents of the documents D2, D3, D5 and D7 include a description about installation of an additional component, and thus the categorization of these documents into the cluster can be regarded as proper.

The reason why this technique allows documents to be properly categorized is that feature elements are first extracted from the titles of the respective documents, and then the documents are categorized on the basis of the extracted feature elements. That is, in most cases, the titles of documents represent, in a simplified fashion, what is described in the contents of the documents. Therefore, if categorization is performed using feature elements included in the titles of documents, scattering into a large number of clusters can be prevented, and the probability of generation of noise clusters is reduced. Furthermore, because the titles are created by the authors of the documents so as to shortly represent what is described in the documents, categorization on the basis of the author's viewpoints can be obtained.

After completion of categorization, if a user issues a command to select a cluster of "paper", documents D1, D4, D6, and D7 categorized in that cluster are read from the document storage unit 11 and displayed on the display unit 32. Herein, only the document numbers or document names may be displayed, or otherwise the contents of the documents may be displayed.

In the present invention, after the clustering process described above, the cluster merging unit 2 performs a cluster merging process.

That is, in the categorization result shown in Fig. 5, the cluster of "paper" includes documents D1, D4, D6, and D7, and the cluster of "cassette" includes documents D1, D4, and D7.

Thus, documents D1, D4, and D7 are included in both clusters of "paper" and "cassette". This means that a feature element of "paper" and a feature element of "cassette" have a close relation with each other. For example, an expression of "paper cassette" is included in the title or the main body of the documents D1, D4, and D7, and thus these documents D1, D4, and D7 can be regarded as having a close relation. Therefore, it is more desirable that these documents D1, D4, and D7 be categorized into the same cluster.

In the present invention, to the above end, after performing the clustering on the basis of the feature elements, the cluster merging process is performed upon the result of the clustering.

The cluster merging process is described below. First, aside from the categorization result shown in Fig. 5, a general example is described with reference to Fig. 6.

We assume here that there are two clusters C1 and C2, wherein the cluster C1 includes five documents D1, D2, D3, D4, and D8, and the cluster C2 includes six documents D3, D4, D5, D6, D7, and D8.

Documents which are commonly included in both clusters C1 and C2 are D3, D4, and D8. In the present embodiment, the degree of relation among a plurality of clusters is evaluated on the basis of the number of documents which are commonly included in the plurality of clusters, and clusters are merged depending upon the evaluated degree of relation.

More specifically, the ratio of the number of documents which are commonly included in two certain clusters to the total number of documents included in those two clusters is calculated, and a decision as to whether these two clusters should be merged is made depending upon whether the calculated ratio is equal to or greater than a predetermined threshold value.

In this specific example, the total number of documents included in the two clusters C1 and C2 is eleven, and three documents are commonly included in both clusters. Thus, the ratio (%) of the number of common documents to the total number of documents can be calculated, and the decision as to whether merging should be performed is made in accordance with the calculation result. In the calculation of the ratio (%), the ratio may be determined simply by dividing the number of common documents by the total number of documents and further multiplying the result by 100, or the ratio may be determined by dividing the product of the number of common documents and a predetermined arbitrary factor by the total number of documents and then multiplying the result by 100.

As an example, let us assume that the number of documents included in the cluster C1 is equal to $\alpha1$, the number of documents included in the cluster C2 is equal to $\alpha2$, and the number of documents which are commonly included in both clusters C1 and C2 is equal to β. After multiplying β by a factor of, for example, 2, $2\beta/(\alpha1 + \alpha2) \times 100$ is calculated. The result (%) is compared with a predetermined threshold value TH (%). If the calculated result is equal to or greater than the threshold value TH, then merging is performed. In the example shown in Fig. 6, $2\beta = 2 \times 3 = 6$, and $\alpha1 + \alpha2 = 5 + 6 = 11$, and thus the ratio is calculated as 55%. If the threshold value TH is set to 70%, the calculated ratio (55%) is smaller than the threshold value TH (70%), and thus it is determined that the clusters C1 and C2 should not be merged. The above factor may be set to an arbitrary value such that the calculated ratio (%) falls within a range which is proper for comparison with the threshold value. Thus, although the factor is set to 2 in the above example, the factor may be set to 1.

Referring back to the categorization result shown in Fig. 5, the cluster of "paper" includes four documents D1, D4, D6, and D7, and the cluster of "cassette" includes three documents D1, D4, and D7. Three documents D1, D4, and D7 are commonly included in both clusters. Now, we calculate the ratio (%) of the number of common documents to the total number of documents.

The calculation is performed in accordance with the formula described above. In the case of the categorization result shown in Fig. 5, the total number of documents ($\alpha1 + \alpha2$) is calculated as 4 + 3 = 7, the number of common documents is equal to 3, and thus 2β is calculated as 6. In this case, the ratio becomes as high as about 86%. Because the calculated ratio is greater than the predetermined threshold value (70% in this example), it is determined that the cluster of "paper" and the cluster of "cassette" should be merged into a single cluster.

Similarly, decisions as to whether the cluster of "paper" and the cluster of "installation" should be merged and whether the cluster of "cassette" and the cluster of "installation" should be merged are made as follows.

As for the clusters of "paper" and "installation", the cluster of "paper" includes four documents D1, D4, D6 and D7, and the cluster of "installation" includes four documents D2, D3, D5, and D7. Only one document D7 is commonly included in both clusters. Thus, according to the formula described above, the ratio is calculated as 25%, which is lower than the threshold value (70%). Therefore, it is determined that these clusters should not be merged.

As for the clusters of "cassette" and "installation", the cluster of "cassette" includes three documents D1, D4, and D7, the cluster of "installation" includes four documents D2, D3, D5, and D7, and only one document D7 is commonly included in both clusters. Thus, according to the formula described above, the ratio is calculated as 28%, which is also lower than the threshold value (70%). Therefore, it is determined that these clusters should not be merged.

As described above, it is determined whether merging should be performed for each combination of two clusters. The result of categorization (recategorization by merging) performed upon the categorization result shown in Fig. 5 is shown in Fig. 7. In Fig. 7, the cluster of "paper" and the cluster of "cassette" are combined into a single cluster of "paper + cassette" including documents D1, D4, D6, and D7. On the other hand, the cluster of "installation" remains in the original state without being combined with another cluster.

Referring to Fig. 7, the recategorization result obtained through the cluster merging process indicates that in the cluster "paper + cassette" (including documents D1, D4, D6, and D7), as can be understood from the document contents shown in Fig. 2, a paper cassette is described in the content of the document D1, setting of paper is described in the document D4, a method of handling which should be performed when printed paper becomes dirty is described in the document D6, and installation of an additional paper cassette is described in the document D7.

The contents of the documents D1, D4, D6 and D7 include a description about paper or a cassette, and thus the recategorization of these documents into the single cluster can be regarded as proper. As a matter of fact, merging into the single cluster of "paper + cassette" results in better categorization.

As described above, a better result can be obtained by first extracting feature elements from the titles of the respective documents, then performing the clustering process on the basis of the extracted feature elements, and finally performing the cluster merging process for each combination of two clusters of the clusters obtained via the above clustering process.

After the first cluster merging process for each combination of two clusters, the result of the recategorization by the cluster merging is obtained as shown in Fig. 7. Thereafter, a second cluster merging process is performed upon the recategorization result obtained through the first cluster merging process. That is, in the result of the first cluster merging process shown in Fig. 7, a cluster merging process is performed for a combination of the cluster of "paper + cassette" and the cluster of "installation". In this example, as for the combination of clusters of "paper + cassette" and "installation", the cluster "paper + cassette" includes four documents D1, D4, D6, the cluster of "installation" includes four document D2, D3, D5, and D7, and only one document D7 is included in both clusters. The ratio (%) of the number of common documents to the total number of documents is calculated as follows. The number of common document, equal to 1, is first multiplied by a factor of 2 and then divided by the total number of documents, equal to 8, and further multiplied by 100, and thus the result is obtained as 25%, which is lower than the threshold value (70%). Thus, it is determined that these clusters should not be merged.

As described above, after completion of the first cluster merging process for each combination of two clusters, the second cluster merging process is performed for each combination of two clusters of the clusters obtained via the fist cluster merging process. After completion of the second cluster merging process, a third cluster merging process is performed for each combination of two clusters of the clusters obtained via the second cluster merging process. The above process is performed repeatedly until no new cluster is created (until no clusters are merged).

Although in the above example, the cluster merging process is performed for a combination of two clusters, the cluster merging process may be performed for a combination of three or more clusters. In this case, in a first cluster merging process, cluster merging is performed for each combination of three or more clusters. Thereafter, cluster merging may be performed repeatedly for the result obtained via the previous cluster merging process until no further merging occurs. As in the previous case, the judgment as to whether three or more clusters should be merged can be performed on the basis of the ratio (%) of the number of common documents to the total number of documents included in these clusters.

In the above-described cluster merging process for combinations of a plurality of clusters, the ratio of the number of documents commonly included in the clusters to the total number of documents is calculated from the categorization result such as that shown in Fig. 5, and the ratio is compared with the predetermined threshold value. Alternatively, the judgment as to whether clusters should be merged or not can be made by examining in what manner feature elements characterizing the respective clusters are used in the original documents. An example of an apparatus for performing the cluster merging process in the above-described manner is shown in Fig. 8. The apparatus shown in Fig. 8 is includes the same constituent parts as those shown in Fig. 1 and they are denoted by the same reference numerals. However, the difference is in that the output of the document storage unit 11 is applied to the cluster merging unit 2 so that the decision as to whether cluster merging should be preformed can be made on the basis of the document contents, as will be described below.

Herein, let us assume that the cluster merging process is performed for the clusters of "paper" and "cassette" shown in Fig. 5. The cluster of "paper" includes documents D1, D4, D6, and D7, and the cluster of "cassette" includes documents D1, D4, and D7.

These documents are examined to detect in what manner the words "paper" and "cassette" are used in the documents. In the document D1, a phrase "paper cassette" which is a combination of "paper" and "cassette" appears at plural locations. The document D4 also includes a phrase "paper cassette". Furthermore, in the document D4, "paper" and "cassette" appear at close locations. The document D7 also includes a phrase "paper cassette" and a phrase "paper cassette unit". Although the document D6 does not include a word "cassette", a word "paper" appears at plural locations.

Form the above, it can be concluded that words "paper" and "cassette" extracted as feature elements are used at adjacent or close locations, and thus they can be regarded as having a close relation. Thus, at least documents D1, D4, and D7 have a close relation, and the document D6 has a relation to some extent. Therefore, it can be concluded that the clusters of "paper" and "cassette" can be combined properly into a single cluster of "paper + cassette".

Thereafter, the cluster merging process is performed for the clusters of "paper" and "installation". The cluster of "paper" includes documents D1, D4, D6, and D7, and the cluster of "installation" includes documents D2, D3, D5, and D7.

These documents are examined to detect in what manner the words "paper" and "cassette" are used in the documents. In the documents D1, D2, D3, D4, D5, and D6, "paper" and "installation" do not appear at adjacent or close locations. Only in the document D7, "paper cassette" and "installation" appear at close locations.

Therefore, it can be concluded that "paper" and "installation" extracted as feature elements are not frequently used at adjacent or close locations, and thus they can be regarded as having little relation. Thus, it is determined that the clusters of "paper" and "installation" should not be merged.

In the case of the cluster merging process for a combination of the clusters of "cassette" and "installation", as in the case of the combination of the clusters of "paper" and "installation", "cassette" and "installation" are not used at adjacent or close locations.

Therefore, it can be concluded that "cassette" and "installation" extracted as "feature elements" are not frequently used at adjacent or close locations, and thus they can be regarded as having little relation. Thus, it is determined that the clusters of "cassette" and "installation" should not be merged.

Also in the case where cluster merging is performed in the above-described manner depending upon in what manner feature elements characterizing the respective clusters are used in the original documents, after completion of a first cluster merging process for each combination of clusters, a second cluster merging process is performed for each combination of two clusters of the clusters obtained via the fist cluster merging process. After completion of the second cluster merging process, a third cluster merging process is performed for each combination of two clusters of the clusters obtained via the second cluster merging process. The above process is performed repeatedly until no further cluster is created (until no clusters are merged).

Although in the above example, the cluster merging process is performed for a combination of two clusters, the cluster merging process may be performed for a combination of three or more clusters. In this case, in a first cluster merging process, cluster merging is performed for each combination of three or more clusters. Thereafter, cluster merging may be performed repeatedly for the result obtained via the previous cluster merging process until no further merging occurs.

It is desirable that when the result obtained via the cluster merging process is presented to a user, information representing how the cluster merging process has been performed be also presented together with the above result to the user. This can be achieved if the information representing in what manner the cluster merging process has been performed by the cluster merging unit 2 is supplied to the output control unit 31, and the output control unit 31 displays the received information on the display unit 32.

Note that the present embodiment is not limited to the specific examples described above, but various modifications are possible without departing from the spirit of the embodiment. For example, although in the above example, feature elements to be used to obtain a categorization result such as that shown in Fig. 5 are extracted from the titles of the respective documents, and clustering is performed on the basis of the feature elements extracted from the titles, what is essential to the present embodiment is that after categorizing documents into clusters according to semantic similarity, similar clusters are merged. Therefore, the manner of clustering a plurality of documents is not limited to a particular method. For example, instead of clustering documents on the basis of feature elements extracted from the titles of the documents according to the above embodiment, clustering may also be performed according to URL addresses (after removing "http://", the remaining part is employed), updated date/time (without any restriction or within last one moth), or file sizes (the sizes of the Web pages in bytes). One of these items or some combination of these items may be employed in the clustering process. A desired item can be selected, for example, from a menu. In the case where a selected item is not included in a document, another item may be employed instead of the selected item. For example, when the title is selected as the item, if a Web page does not include a title, a URL address may be employed.

After performing the clustering using one of the methods, the judgment as to whether clusters should be merged or not is made by evaluating the similarity between the clusters under consideration in the manner described above.

For example, let us consider an example in which clustering is performed according to URL addresses. We assume herein that documents have been categorized into a cluster of a certain URL (URL1) and a cluster of another URL (URL1). We further assume that the cluster of URL1 includes documents D1, D2, D3, and D4, and the cluster of URL2 includes documents D2, D3, D4, D5. In this case, documents which are commonly included in both clusters are D2, D3, and D4. The ratio of the number of common documents to the total number of documents is calculated, and it is determined according to the calculated ratio whether the cluster of URL1 and the cluster of URL2 should be merged or not.

Although in the above embodiment, the judgment as to whether clusters should be merged or not is made by comparing the ratio (%) of the number of documents commonly included in clusters under consideration to the total number of documents to a predetermined threshold value (%), the manner of the judgment is not limited to that. For example, the judgment as to whether clusters should be merged or not may be made in accordance with the number of common documents relative to the numbers of documents included in the respective clusters.

In the above embodiment, different documents D1, D2,..., D7 are categorized. The embodiment may also be applied to the case where a single document is divided into a plurality of contents (into parts having their own themes) and the respective contents are categorized. Herein, let us assume that contents are given by dividing a single document at respective titles into plural parts each describing their own particular themes.

For example, if it is assumed that the documents D1, D2,..., D7 shown in Fig. 7 are parts of the same single document, these documents D1, D2,...., D7 can be regarded as contents in the above-described sense. In this case, the respective contents include titles T1, T2,..., T7 and main bodies A1, A2,..., A7.

As described above, the present embodiment may be applied to the case when a signal document is divided into a plurality contents, then the contents are categorized into clusters, and finally similar clusters in the obtained clusters are merged.

Furthermore, the present embodiment may also be applied to a plurality of documents obtained via a general retrieval service. In this case, the clustering process is first performed for a large number of documents obtained via the retrieval, and the cluster merging process is then performed for the result of the clustering process.

A program used to execute the above-described document categorizing process according to the present embodiment may be stored on a storage medium such as a floppy disk, an optical disk, or a hard disk. Note that such a storage medium also falls within the scope of the present invention. The program may also be obtained via a network.

(Second Embodiment)

When merged clusters are presented to a user, if only the final result of the cluster merging process is presented and no information about the cluster merging process is presented, the user cannot know which clusters have been combined together into final clusters and cannot know the degree of relation among the original clusters combined together into the final clusters.

In the present embodiment, to solve the above problem, when final clusters, which are obtained through the cluster merging process in which clusters having close relations are combined together, are displayed, the clusters are displayed in a manner that allows a user to see which clusters have been combined together into which final clusters and also see the degrees of relation among the clusters combined together.

The second embodiment of the present invention is described in further detail below.

In this second embodiment, categorization of documents is performed, as described above, by first extracting the titles of the respective documents, then extracting feature elements from the titles, and finally categorizing the documents on the basis of the extracted feature elements.

Fig. 9 illustrates the second embodiment. As shown in Fig. 9, an apparatus of the second embodiment consists mainly of a clustering unit 91 for categorizing documents into a plurality of clusters in accordance with semantic similarity, a cluster merging unit 92 which evaluates the relation among the plurality of clusters created by the clustering unit 91 on the basis of the documents included in the respective clusters and then combines two or more clusters having a degree of relation equal to or higher than a predetermined value, a cluster-merging-process information generator 93 for generating cluster-merging-process information representing which clusters have been combined together and also representing the degrees of relations among the combined clusters wherein the cluster-merging-process information is to be displayed when final clusters obtained via the cluster merging process performed by the cluster merging unit 2 are displayed, and a categorization result outputting unit 94 for outputting the categorization result including the cluster-merging-process information.

The clustering unit 91 includes a document storage unit 911, a sentence analyzer 912, a feature element extractor 913, a feature table generator 914, a document categorizing unit 915, and a categorization result storage unit 16.

The document storage unit 911 stores, in the form of a database, a large number of document data. Hereinafter, it is assumed that the documents shown in Fig. 10 are categorized. The documents shown in Fig. 10 include different documents D1, D2,..., D7 including titles T2, T2,..., T7 and main bodies A1, A1,..., A7, respectively.

The sentence analyzer 912 analyzes the documents stored in the document storage unit 911 to extract the titles of the respective documents. The extraction of the titles is performed by the document analyzer 912 as follows.

A first method is to detect a part defined as a title according to the document format, and employ the detected part as a title if such a part is detected. A second method is to detect a part specified to be displayed with characters having a greater size than a standard size, and employ the detected part as a title if such a part is detected. A third method is to extract a predetermined number of sentences or words located at the beginning of a document and employ the extracted sentence or words as a title. The first, second, and third methods described above are performed sequentially. When the first method is performed, if a part defined as a title is detected, the detected part is employed as the title. If a part defined as a title is not detected, the second method is performed. If a part specified to be displayed with large-sized characters is detected, the detected part is employed as the title. If a part specified to be displayed with large-sized characters is not detected, the third method is performed to detect a title.

The feature element extractor 913 extracts a feature element from the respective document titles detected by the sentence analyzer 2.

The feature table generating means 914 produces a feature table representing the relationship between the feature elements detected from the titles and the respective documents. A specific content of the feature table will be described later.

The document categorizing unit 915 examines the content of the feature table and categorizes the documents D1, D2,..., D7 into a plurality of clusters according to semantic similarity. Documents including a common feature element are detected on the basis of the feature elements included in the titles of the documents D1, D2,..., D7, and the detected documents are categorized into a cluster. The document categorization unit 915 includes a synonymous feature dictionary (not shown). Categorization into clusters may also be performed using the synonymous feature dictionary as follows. When documents including a common feature element are detected, a judgment as to whether a common feature element is included or not is made by determining whether a synonym is included or not using the synonymous feature dictionary. If synonyms are detected in documents, those documents are categorized into the same cluster.

The categorization result storage unit 916 stores the content obtained via the categorization performed by the document categorization unit 915.

The cluster merging unit 92 evaluates the relation among the plurality of clusters on the basis of the documents included in the respective clusters and then combines two or more clusters the degree of relation among which is equal to or higher than a predetermined value, as will be described in detail later.

The cluster-merging-process information generator 93 includes a relation evaluator 931 and a manner-of-displaying-cluster-name determiner 932, wherein the relation evaluator 931 evaluates the degree of relation among clusters by comparing a cluster correlation score (described later) generated by the cluster merging unit 92 with a predetermined threshold value (described later), and the manner-of-displaying-cluster-name determiner 932 determines the manner of displaying cluster names so as to indicate which clusters have been combined together and indicate the degree of relation among the combined clusters, on the basis of the degree of relation evaluated by the relation evaluator 931. The processing performed by the relation evaluator 931 and the manner-of-displaying-cluster-name determiner 932 will be described in further detail later.

The categorization result output unit 94 includes an output control unit 941 and a display unit 942 and serves to output the document categorization result obtained according to the present invention.

The document categorization process performed by the apparatus constructed in the above-described manner according to the present invention is described below. The outline of the document categorization process according to the present embodiment is as follows. As shown in a flow chart in Fig. 11, a large number of documents to be processed are first categorized into a plurality of clusters according to semantic similarity (step 11S1). Thereafter, the degrees of relation among clusters are evaluated on the basis of the documents included in the respective clusters (step 11S2). Two or more clusters having a degree of relation higher than a predetermined value are combined together (step 11S3). Thereafter, cluster-merging-process information is generated which indicates which clusters have been merged into final clusters and also indicates the degrees of relation among the original clusters combined together. More specifically, the degrees of relation among the clusters which have been merged are determined (step 11S4), and cluster-merging-process information is generated on the basis of the degrees of relation so that the cluster-merging-process information represents the properties of the original clusters combined together into the final clusters, that is, so that the cluster-merging-process information indicates which clusters have been combined together into final clusters and also indicates the degrees of relation among the original clusters combined together (step 11S5). The document categorization process is described in further detail below with reference to a specific example.

Herein, it is assumed that the documents D1, D2,..., D7 shown in Fig. 10 are categorized. In the present embodiment, feature elements are extracted from the titles of the respective documents, and clustering is performed on the basis of the extracted feature elements. Thereafter, obtained clusters are merged. First, the process of extracting feature elements from the titles and performing the clustering on the basis of the extracted feature elements (by the clustering unit 1) is described.

The document analyzer 12 detects the titles of the respective documents D1, D2,..., D7. For example, the title T1 is detected from the document D1, the title T2 from the document D2, the title T3 from the document D3, and so on. Thus, the titles T1, T2,..., T7 are detected from the respective documents D1, D2,..., D7.

The feature element extractor 913 then extracts feature elements from the respective titles. Thereafter, the feature table generator 914 produces a feature table representing the relationships between the feature elements and the documents including the feature element in their titles. An example of a feature table is shown in Fig. 12. In this example, the feature table represents the relationships between feature elements each included in three or more different documents and the documents including the feature elements. Numerals described in the feature table represent the numbers of feature elements included in the respective documents. For example, in the case of feature element "paper", one feature element is included in the title of each of the documents D1, D4, D6, and D7.

As can be seen from the feature table shown in Fig. 12, documents including "paper" as a feature element in their title are D1, D4, D6, and D7, documents including "cassette" as a feature element in their title are D1, D4, and D7, and documents including "installation" as a feature element in their title are D2, D3, D5 and D7. In Fig. 10, these feature elements are underlined.

The document analyzer 915 access the feature table and categorizes the respective documents into clusters for each feature element. The result of the categorization is shown in Fig. 13. In the categorization into clusters, as described earlier, the detection of feature elements commonly included in documents may be performed by detecting synonyms included in documents, using the synonymous feature dictionary, and documents including detected synonyms may be categorized into the same document cluster. For example, when "paper" and "printing paper" are extracted as feature elements, documents including either one of these feature elements are categorized into the same cluster.

The obtained categorization result is stored in the categorization result storage unit 916. In the categorization result shown in Fig. 13, as for a cluster categorized as "paper" (including documents D1, D4, D6, and D7), as can be understood from the document contents shown in Fig. 10, a paper cassette is described in the content of the document D1, setting of paper in the document D4, a smear created on printed paper in the document D6, and installation of an additional paper cassette in the document D7.

Thus, the categorization of the documents D1, D4, D6, and D7 including a description about paper into the cluster can be regarded as proper.

In the case of a cluster categorized as "cassette" (including documents D1, D4, and D7), as can be understood from the document contents shown in Fig. 10, a paper cassette is described in the content of the document D1, setting of paper in the document D4, and installation of an additional paper cassette in the document D7.

The contents of the documents D1, D4, D6 and D7 include a description about setting of paper, and thus the categorization of these documents into the cluster can be regarded as proper.

In the case of a cluster categorized as "installation" (including documents D2, D3, D5, and D7), as can be understood from the document contents shown in Fig. 10, installation of extension memory is described in the content of the document D2, installation of an interface card in the document D3, installation of an additional hard disk in the document D5, and installation of an additional paper cassette in the document D7.

The contents of the documents D2, D3, D5 and D7 include a description about installation of an additional part, and thus the categorization of these documents into the cluster can be regarded as proper.

The reason why this technique allows documents to be properly categorized is that feature elements are first extracted from the titles of the respective documents, and then the documents are categorized on the basis of the extracted feature elements. That is, in most cases, the titles of documents represent, in a simplified fashion, what is described in the contents of the documents. Therefore, if categorization is performed using feature elements included in the titles of documents, scattering into a large number of clusters can be prevented, and the probability of generation of noise clusters is reduced. Furthermore, because the titles are created by the authors of the documents so as to shortly represent what is described in the documents, categorization on the basis of the author's viewpoints can be obtained.

After completion of categorization, if a user issues a command to select a cluster of "paper", documents D1, D4, D6, and D7 categorized in that cluster are read from the document storage unit 11 and displayed on the display unit 32. Herein, only the document numbers or document names may be displayed, or otherwise the contents of the documents may be displayed.

In the present invention, after the clustering described above, the cluster merging unit 2 performs a cluster merging process.

That is, in the categorization result shown in Fig. 13, the cluster of "paper" includes documents D1, D4, D6, and D7, and the cluster of "cassette" includes documents D1, D4, and D7.

Thus, documents D1, D4, and D7 are included in both clusters of "paper" and "cassette". This means that a feature element of "paper" and a feature element of "cassette" have a close relation with each other. For example, an expression of "paper cassette" is included in the title or the main body of the documents D1, D4, and D7, and thus these documents D1, D4, and D7 can be regarded as having a close relation. Therefore, it is more desirable that these documents D1, D4, and D7 be categorized into the same cluster.

To the above end, after performing the clustering on the basis of the feature elements, the cluster merging process is performed upon the result of the clustering.

The cluster merging process is described below. First, aside from the categorization result shown in Fig. 13, a general example is described with reference to Fig. 14.

We assume here that there are two clusters C1 and C2, wherein the cluster C1 includes five documents D1, D2, D3, D4, and D8, and the cluster C2 includes six documents D3, D4, D5, D6, D7, and D8.

Documents which are commonly included in both clusters C1 and C2 are D3, D4, and D8. In the present embodiment, the degree of relation among a plurality of clusters is evaluated on the basis of the number of documents which are commonly included in the plurality of clusters, and clusters are merged depending upon the evaluated degree of relation.

More specifically, the ratio of the number of documents which are commonly included in two certain clusters to the total number of documents included in those two clusters is calculated, and a decision as to whether these two clusters should be merged is made depending upon whether the calculated ratio is equal to or greater than a predetermined threshold value.

In this specific example, the total number of documents included in the two clusters C1 and C2 is eleven, and three documents are commonly included in both clusters. Thus, the ratio (%) of the number of common documents to the total number of documents can be calculated, and the decision as to whether merging should be performed is made in accordance with the calculation result. When the ratio (%) is calculated, the ratio may be determined simply by diving the number of common documents by the total number of documents and further multiplying the result by 100, or the ratio may be determined by dividing the product the number of common documents and a predetermined arbitrary factor by the total number of documents and then multiplying the result by 100.

As an example, let us assume that the number of documents included in the cluster C1 is equal to α1, the number of documents included in the cluster C2 is equal to α2, and the number of documents which are commonly included in both clusters C1 and C2 is equal to β. After multiplying β by a factor of, for example, 2, and then 2β/(α1 + α2) × 100 is calculated. The result (%) is compared with a predetermined threshold value TH (%). If the calculated result is equal to or greater than the threshold value TH, then merging is performed. In the example shown in Fig. 14, 2β = 2 × 3 = 6, and α1 + α2 = 5 + 6 = 11, and thus the ratio is calculated as 55%. If the threshold value TH is set to 70%, the calculated ratio (55%) is smaller than the threshold value TH (70%), and thus it is determined that the clusters C1 and C2 should not be merged. The above factor may be set to an arbitrary value such that the calculated ratio (%) falls within a range which is proper for comparison with the threshold value. Thus, although the factor is set to 2 in the above example, the factor may be set to 1.

Referring back to the categorization result shown in Fig. 13, In this example, the cluster of "paper" includes four documents D1, D4, D6, and D7, and the cluster of "cassette" includes three documents D1, D4, and D7. Three documents D1, D4, and D7 are commonly included in both clusters. Now, we calculate the ratio (%) of the number of common documents to the total number of documents.

The calculation is performed in accordance with the formula described above. In the case of the categorization result shown in Fig. 13, the total number of documents (α1 + α2) is calculated as 4 + 3 = 7, the number of common documents is equal to 3, and thus 2β is calculated as 6. In this case, the ratio becomes as high as about 86%. Because the calculated ratio is greater than the predetermined threshold value (70% in this example), it is determined that the cluster of "paper" and the cluster of "cassette" should be merged into a single cluster.

Similarly, decisions as to whether the cluster of "paper" and the cluster of "installation" shown in Fig. 13 should be merged and whether the cluster of "cassette" and the cluster of "installation" should be merged are made as follows.

As for the clusters of "paper" and "installation", the cluster of "paper" includes four documents D1, D4, D6, and D7, the cluster of "installation" includes four documents D2, D3, D5, and D7, and only one document D7 is included in both clusters. Thus, according to the formula described above, the ratio is calculated as 25%, which is lower than the threshold value (70%). Therefore, it is determined that these clusters should not be merged.

As for the clusters of "cassette" and "installation", the cluster of "cassette" includes three documents D1, D4, and D7, the cluster of "installation" includes four documents D2, D3, D5, and D7, and only one document D7 is included in both clusters. Thus, according to the formula described above, the ratio is calculated as 28%, which is also lower than the threshold value (70%). Therefore, it is determined that these clusters should not be merged.

As described above, it is determined whether merging should be performed for each combination of two clusters. The result of categorization (recategorization by merging) performed upon the categorization result shown in Fig. 13 is shown in Fig. 15. In Fig. 15, the cluster of "paper" and the cluster of "cassette" are combined into a single cluster of "paper + cassette" including documents D1, D4, D6, and D7. On the other hand, the cluster of "installation" remains in the original state without being combined with another cluster.

Referring to Fig. 15, the recategorization result obtained through the cluster merging process indicates that in the cluster "paper + cassette" (including documents D1, D4, D6, and D7), as can be understood from the document contents shown in Fig. 10, a paper cassette is described in the content of the document D1, setting of paper in the document D4, a method of handling which should be performed when printed paper becomes dirty is described in the document D6, and installation of an additional paper cassette in the document D7.

The contents of the documents D1, D4, D6 and D7 include a description about paper or a cassette, and thus the recategorization of these documents into the single cluster can be regarded as proper. As a matter of fact, merging into the single cluster of "paper + cassette" results in better categorization.

As described above, a better result can be obtained by first extracting feature elements from the titles of the respective documents, and clustering is performed on the basis of the extracted feature elements, and finally performing the cluster merging process for each combination of two clusters of the clusters obtained via the above clustering process.

After the first cluster merging process for each combination of two clusters, the result of the recategorization by the cluster merging is obtained as shown in Fig. 15. Thereafter, a second cluster merging process is performed upon the recategorization result obtained through the first cluster merging process. That is, in the result of the first cluster merging process shown in Fig. 15, a cluster merging process is performed for a combination of the cluster of "paper + cassette" and the cluster of "installation". In this example, as for the combination of clusters of "paper + cassette" and "installation", the cluster "paper + cassette" includes four documents D1, D4, D6, the cluster of "installation" includes four document D2, D3, D5, and D7, and only one document D7 is included in both clusters. The ratio (%) of the number of common documents to the total number of documents is calculated as follows. The number of common document, equal to 1, is first multiplied by a factor of 2 and then divided by the total number of documents, equal to 8, and further multiplied by 100, and thus the result is obtained as 25%, which is lower than the threshold value (70%). Thus, it is determined that these clusters should not be merged.

After completion of the first cluster merging process for each combination of two clusters, a second cluster merging process is performed for each combination of two clusters of the clusters obtained via the fist cluster merging process. After completion of the second cluster merging process, a third cluster merging process is performed for each combination of two clusters of the clusters obtained via the second cluster merging process. The above process is performed repeatedly until no further cluster is created (until no clusters are merged).

Although in the above example, the cluster merging process is performed for a combination of two clusters, the cluster merging process may be performed for a combination of three or more clusters. In this case, in a first cluster merging process, cluster merging is performed for each combination of three or more clusters. Thereafter, cluster merging may be performed repeatedly for the result obtained via the previous cluster merging process until no further merging occurs. As in the previous case, the judgment as to whether three or more clusters should be merged can be performed on the basis of the ratio (%) of the number of common documents to the total number of documents included in these clusters.

After the cluster merging unit 92 shown in Fig. 9 completes the cluster merging process, the cluster-merging-process information generator 93 determines the degrees of relation among the original clusters merged together by the cluster merging unit 92 and generates cluster-merging-process information on the basis of the degrees of relation so that the cluster-merging-process information represents the properties of the original clusters combined together into the final clusters, that is, so that the cluster-merging-process information indicates which clusters have been combined together into final clusters and also indicates the degrees of relation among the original clusters combined together. The process performed by the cluster-merging-process information generator 93 is described in further detail below.

In the present embodiment, the relation evaluator 931 evaluates the degrees of relation among clusters merged together by determining whether the cluster correlation scores (%) calculated by the cluster merging unit 92 are much greater than the above-described threshold value TH or close to the threshold value TH. More specifically, a threshold value TH1 is set to a value (%) higher than the above-described threshold value TH, and if the cluster correlation score (denoted by K) calculated by the cluster merging unit 92 is equal to or higher than TH1 (K ≥ TH1, the clusters are determined as having very close relation, that is, as being very similar to each other. On the other hand, if the cluster correlation score K calculated by the cluster merging unit 92 is within a range TH1 > K ≥ TH, the clusters are determined as being similar to each other to a certain extent.

If K ≥ TH1, that is, if clusters merged into a final cluster have been determined as having very close relation, the following process is further performed.

In the case of the specific example shown in Fig. 15, a final cluster created via the cluster merging process has a feature element of "paper + cassette". This cluster of "paper + cassette" is obtained as a result of merging the cluster of "paper" and the cluster of "cassette" shown in Fig. 13.

The clusters may be named as follows. For example, a cluster having a feature element of "paper" is named "paper cluster", and a cluster having a feature element of "cassette" is named "cassette cluster". Hereinafter, the cluster names are represented more simply as "paper" and "cassette".

The cluster correlation score of the cluster of "paper + cassette" created by the cluster merging process has been calculated as 86% by the cluster merging unit 92. Herein we assume that the threshold value TH1 used by the relation evaluator 931 to evaluate the degree of relation is set 80%. In this case, the cluster correlation score K calculated by the cluster merging unit 92 satisfies the condition K ≥ TH1, and thus the paper cluster and the cassette cluster are determined as having very close relation and being very similar to each other.

As described above, when a cluster correlation score K calculated by the cluster merging unit 92 is equal to or greater than TH1 (K ≥ TH1), original clusters merged together into a final cluster can be regarded as having very close relation and being very similar to each other. Thus, in such a case, the name of the final cluster is given by a combination of the original cluster names which are displayed by successively the original cluster names. In the case of "paper cluster" and "cassette cluster", the cluster names "paper" and "cassette" can be combined into "paper cassette".

That is, in this case, the cluster names are displayed in an AND form. This method is employed when a simple combination of cluster names does not result in a problem. In this specific example, the final cluster created via the cluster merging process is named "paper cassette". The naming of the final cluster as "paper cassette" can be judged as proper from the contents of the documents (Fig. 10) in the paper cluster and the cassette clusters merged into the final cluster.

Fig. 16 illustrates an example of information displayed after the above process. In this specific example shown in Fig. 16, "paper cassette" is displayed as the cluster name of the final cluster created via the cluster merging process, and the title names of the documents (D1, D4, D6, D7, shown in Fig. 10) included in this cluster are displayed.

Instead of displaying the original cluster names successively in a single line as shown in Fig. 16, the individual original cluster names "paper" and "cassette" corresponding to the original clusters may be displayed in different adjacent lines as shown in Fig. 17.

In the case where the original cluster names are displayed in different lines, unnatural or incongruous linguistic continuity can be avoided. Although in this specific example, no problems occur when "paper" and "cassette" are combined and represented in a single line as "paper cassette", combining of cluster names into a single line can be incongruous depending upon the specific cluster names. For example, aside from the above example, when a final cluster is created by merging clusters having names of "product", "usage", "outline", if the cluster names is displayed successively in a single line, the result is "product usage outline". Although this is not absolutely unclear in meaning, it is somewhat incongruous in a linguistic sense. In such a case, language processing may be performed to obtain a better expression such as "outline of usage of products". However, the language processing would be complicated and a long processing time would be needed.

In this specific case, the incongruence can be avoided by displaying "product", "usage", and "outline" in different lines. Another advantage of displaying cluster names in different lines is that when a large number of clusters are combined together, displaying of cluster names in different lines prevents the cluster names from extending over a too long length along a horizontal line.

As described above, when the cluster correlation score K calculated by the cluster merging unit 92 satisfies the condition K ≥ TH1, the cluster names of original clusters merged into a final cluster are displayed in the AND form in which the original cluster names are arranged in a single horizontal line or displayed in different lines.

This makes it possible for the user to easily understand which clusters have been merged into a final cluster simply by seeing the cluster name of the final cluster. For example, in the case of the specific example shown in Fig. 16 or Fig. 17, it can be easily understood that the final cluster has been created by combining the original clusters having cluster names "paper" and "cassette" and that the original clusters have very close relation, that is, the documents included therein have similar contents.

When TH1 > K ≥ TH, that is, when the degree of relation among original clusters which have been merged into a final cluster is not very high but some similar documents are included in the clusters, the process is performed as follows.

When the cluster correlation score K calculated by the cluster merging unit 92 is within the range TH1 > K ≥ TH, the original cluster names are represented in an OR form.

In the specific example described above, the original cluster names "product", "usage", and "outline" are represented not in a simple successive fashion but in a fashion in which a delimiter is placed between adjacent cluster names such as ""product·usage·outline". If a user is informed in advance that a delimiter placed between adjacent cluster names represents "OR", the user can understand that the final cluster obtained via the cluster merging process includes some documents having contents about "product", "usage", or "outline". When a final cluster name is represented in the OR form, the delimiter placed between original cluster names is not limited to a dot as is used in "produce·usage·outline" but other types of delimiters may be used. For example, "/" may be placed between original cluster names such as "product/usage/outline".

In some cases, the cluster correlation score K for some original clusters included in a final cluster is equal to or greater than TH1 (K ≥ TH1) but the cluster correlation score K for some other original clusters in the same final cluster is in the range TH1 > K ≥ TH. In this case, the final cluster name is represented in the form of a mixture of AND and OR expressions so that the degrees of relation are indicated in the final cluster name.

In some cases, in a final cluster obtained by merging original clusters, some original clusters may be included in another original cluster. For example, when clusters having cluster names "product", "television" "radio", and "video" are merged together into a final cluster, if the respective clusters "television", "radio" and "video" are included in the cluster "product" and if the cluster correlation scores K are within the range TH1 > K ≥ TH, the cluster name of the final cluster is expressed as "product·(television·radio·video)". The dots in this expression indicate that "product", "television", "radio", and "video" have relation of OR. Furthermore, the brackets enclosing therein "television", "radio", and "video" indicates that clusters "television", "radio", and "video" are included in the cluster "product".

As described above, only by seeing the cluster names of the final clusters obtained via the cluster merging process, it is possible to know which clusters have been combined together into which final clusters and also can know the degrees of relations among the original clusters combined together.

Note that the present embodiment is not limited to the specific examples described above, but various modifications are possible without departing from the spirit of the embodiment. For example, although in the above example, feature elements to be used to obtain a categorization result such as that shown in Fig. 13 are extracted from the titles of the respective documents, and clustering is performed on the basis of the feature elements extracted from the titles, the manner of clustering a plurality of documents is not limited to such a particular method.

For example, instead of clustering documents on the basis of feature elements extracted from the titles of the documents according to the above embodiment, clustering may also be performed according to URL addresses (after removing "http://", the remaining part is employed), updated date/time (without any restriction or within last one moth), or file sizes (the sizes of the Web pages in bytes). One of these items or some combination of these items may be employed in the clustering process. A desired item can be selected, for example, from a menu. In the case where a selected item is not included in a document, another item may be employed instead of the selected item. For example, when the title is selected as the item, if a Web page does not include a title, a URL address may be employed.

After performing the clustering using one of the methods, the judgment as to whether clusters should be merged or not is made by evaluating the similarity between the clusters under consideration in the manner described above.

For example, let us consider an example in which clustering is performed according to URL addresses. We assume herein that documents have been categorized into a cluster of a certain URL (URL1) and a cluster of another URL (URL1). We further assume that the cluster of URL1 includes documents D1, D2, D3, and D4, and the cluster of URL2 includes documents D2, D3, D4, and D5. In this case, documents which are commonly included in both clusters are D2, D3, and

D4. The ratio of the number of common documents to the total number of documents is calculated, and it is determined according to the calculated ratio whether the cluster of URL1 and the cluster of URL2 should be merged or not.

Although in the above embodiment, the judgment as to whether clusters should be merged or not is made by comparing the ratio (%) of the number of documents commonly included in clusters under consideration to the total number of documents to a predetermined threshold value (%), the manner of the judgment is not limited to that. For example, the judgment as to whether clusters should be merged or not may be made in accordance with the number of common documents relative to the numbers of documents included in the respective clusters.

When the judgment as to whether to merge clusters is made on the basis of the number of clusters, the threshold value may be represented in the number of clusters. For example, when the total number of document is 10, if it is desired to perform merging when the number of common documents is equal to greater than 7, then the threshold value TH is set to 7 and TH1 to 9. In this case, when the number of common documents is equal to or greater than 9, the cluster name of a resultant cluster obtained via the cluster merging process is expressed in the AND form, while when the number of common documents is within the range from 7 to 8, the cluster name of a resultant cluster is expressed in the OR form. Note that the above threshold values used herein or in the previous embodiment are given as mere examples, and they are not limited to those specific examples.

In the above embodiment, different documents D1, D2,..., D7 are categorized. The embodiment may also be applied to the case where a single document is divided into a plurality of contents (into parts having their own themes) and the respective contents are categorized. Herein, let us assume that contents are given by dividing a single document at respective titles into plural parts each describing their own particular themes.

For example, if it is assumed that the documents D1, D2,..., D7 shown in Fig. 10 are parts of the same single document, these documents D1, D2,...., D7 can be regarded as contents in the above-described sense. In this case, the respective contents include titles T1, T2,..., T7 and main bodies A1, A2,..., A7.

As described above, the present invention may be applied to the case when a signal document is divided into a plurality contents, then the contents are categorized into clusters, and finally similar clusters in the obtained clusters are merged.

Furthermore, the present embodiment may also be applied to a plurality of documents obtained via a general retrieval service. In this case, the clustering process is first performed for a large number of documents obtained via the retrieval, and the cluster merging process is then performed for the result of the clustering process. Thereafter, the process described above may be performed upon the clusters obtained via the cluster merging process so that it can be easily see which clusters have been merged together into which clusters and see the degrees of the relations.

A program used to execute the above-described document categorizing process according to the present embodiment may be stored on a storage medium such as a floppy disk, an optical disk, or a hard disk. Note that such a storage medium also falls within the scope of the present invention. The program may also be obtained via a network.